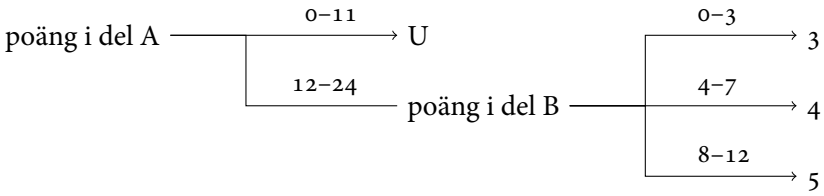
TDDD01 Språkteknologi (2015)

Tentamen 2015-06-08

Marco Kuhlmann

Denna tentamen består av två delar: del A, som innehåller uppgifter 1–8, och del B, som innehåller uppgifter 9–12. Varje uppgift är värd 3 poäng. Betyget sätts enligt följande schema:



Dina inlämningar till del B kommer endast att rättas om du har minst 12 poäng i del A. Rättningen av del A kommer då att avbrytas.

Lycka till!

Del A

1. Ett textklassificeringssystem baserat på metoden Naive Bayes klassificerar nyhetstexter som antingen ”nyheter om Sverige” (S) eller ”nyheter om Norge” (N). Systemet använder följande sannolikheter:

   $$P(\mathtt{S}) = 3/4 \qquad P(\text{Stockholm} \mid \mathtt{S}) = 5/8 \qquad P(\text{Oslo} \mid \mathtt{S}) = 1/8$$
   $$P(\mathtt{N}) = 1/4 \qquad P(\text{Stockholm} \mid \mathtt{N}) = 1/3 \qquad P(\text{Oslo} \mid \mathtt{N}) = 2/3$$

   (a) Ställ upp bråk för de värden (Π-värden) som systemet jämför för att avgöra om dokumentet ”Stockholm Stockholm Stockholm Oslo” är en nyhet om Sverige eller en nyhet om Norge. (b) Ange en dokumentsamling utifrån vilken man får de angivna sannolikheterna om man skattar med Maximum Likelihood-metoden.

2. Förklara begreppen token, lemma och lexem. Ge för varje begrepp ett exempel på en språkteknologisk tillämpning där begreppet är relevant.

3. I en korpus innehållande 1 215 396 token och 105 436 unika ord hittas ordet *det* 13 694 gånger, ordet *är* 13 700 gånger, ordet *nalkas* 2 gånger, sekvensen *det är* 927 gånger, och sekvensen *det nalkas* 0 gånger.

(a) Ställ upp bråk för ML-skattningen (Maximum Likelihood) av unigramsannolikheten $P(\textit{är})$ och bigramsannolikheten $P(\textit{är} \mid \textit{det})$.

(b) Ställ upp ett bråk för ML-skattningen av bigramsannolikheten $P(\textit{nalkas} \mid \textit{det})$ med Add One-utjämning. Antag att vokabulären består av alla unika ord.

4. Här är två taggsekvenser för meningen *Jag skrev på utan att tveka*:

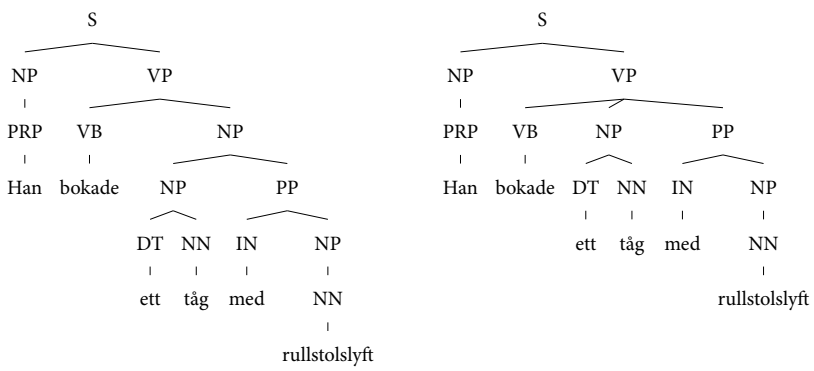
	Jag	skrev	på	utan	att	tveka
sekvens 1	PN	VB	PP	PP	IE	VB
sekvens 2	PN	VB	PL	PP	IE	VB

Du vet redan sannolikheten som en viss Hidden Markov-modell tilldelar sekvens 1 och vill nu veta sannolikheten som modellen tilldelar sekvens 2. Du kan dock endast fråga efter enstaka sannolikheter i modellen och varje sådan fråga kostar 10 kronor. Vilka frågor måste du ställa för att betala så lite som möjligt, och vilken formel måste du använda för att beräkna sekvenssannolikheten?

5. Nedan anges alla NP- och VP-regler i en viss probabilistisk kontextfri grammatik. Den sista NP-regeln och båda VP-regler saknar sannolikhetsvärden.

$\text{NP} \to \text{PRP}\ \frac{2}{7},\ \text{NP} \to \text{NP PP}\ \frac{1}{7},\ \text{NP} \to \text{DT NN}\ \frac{2}{7},\ \text{NP} \to \text{NN},\ \text{VP} \to \text{VB NP},\ \text{VP} \to \text{VB NP PP}$

(a) Vilket sannolikhetsvärde borde den sista NP-regeln ha? (b) Välj sannolikhetsvärden för VP-reglerna sådana att det vänstra trädet nedan får en högre sannolikhet än det högra. Beräkna sannolikheter för båda träd. Antag att alla regler som inte anges ovan har sannolikhet 1. Det räcker att ställa upp bråk.



6. Tre frågor om WordNet: (a) Vad representerar noderna i WordNet? (b) Vilken semantisk relation representerar länkarna? (c) Låt s_1 och s_2 vara två noder i WordNet och låt p stå för den kortaste vägen mellan dessa. Ange en formel för att räkna ut den semantiska likheten mellan s_1 och s_2.

7. Ett namnigenkänningssystem testades på en samling testdata innehållande 800 namnförekomster. Av dessa namn bestod 500 av ett ord, 250 av två ord och 50 av tre ord. Tabellen nedan anger systemets resultat.

	korrekta	falska
Ettordsnamn	420	60
Tvåordsnamn	200	40
Treordsnamn	44	12

Ställ upp bråk för följande:

(a) systemets precision på tvåordsnamn
(b) systemets recall (täckning) på treordsnamn
(c) systemets precision på samtliga namn

8. Olika frågor är olika svåra för automatiska frågebesvarande system att hitta rätt svar på. Rangordna följande tre frågor från lättast till svårast och motivera rangordningen. Systemet antas inte ha någon egen kunskapsbas utan använda sig av Wikipedia-artiklar. Använd relevant terminologi.

A. Vilket år föddes Caesar? B. Hur var han som statsman? C. Var dog han?

Del B

9. Förklara hur CKY-algoritmen för parsning med probabilistiska kontextfria grammatiker fungerar. Vad gör den? På vilken grundidé bygger dess effektivitet?

10. I flera typer av språkteknologiska system kan täckning (recall) inte mätas på det vanligaste sättet, dvs. genom att dividera antalet fall där system och facit överensstämmer med det totala antalet fall i facit. Ange två typer av system där detta inte fungerar så bra, förklara varför, och beskriv de utvärderingsmått som används i stället.

11. Förklara modellen ”den brusiga kanalen” som används i samband med maskinöversättning. Hur kan man skatta de sannolikheter som ingår i denna modell?

12. Hur skulle du kunna använda de språkteknologiska resurser, tekniker och verktyg som du lärt känna under kursen för att skapa ett enkelt system för informationsextraktion från text? Hur skulle du kunna utvärdera systemet?